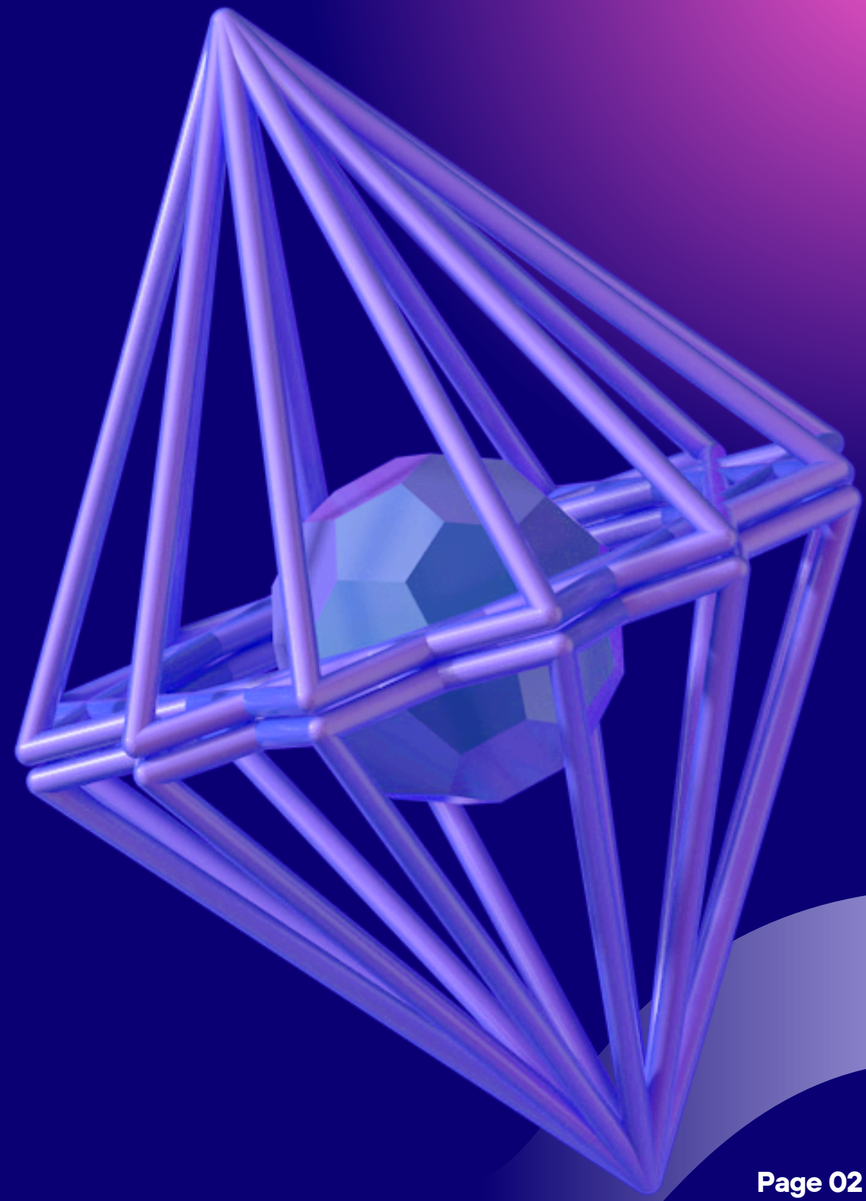
QUERY DATA FROM S3 TO ATHENA WITH AWS GLUE CRAWLER

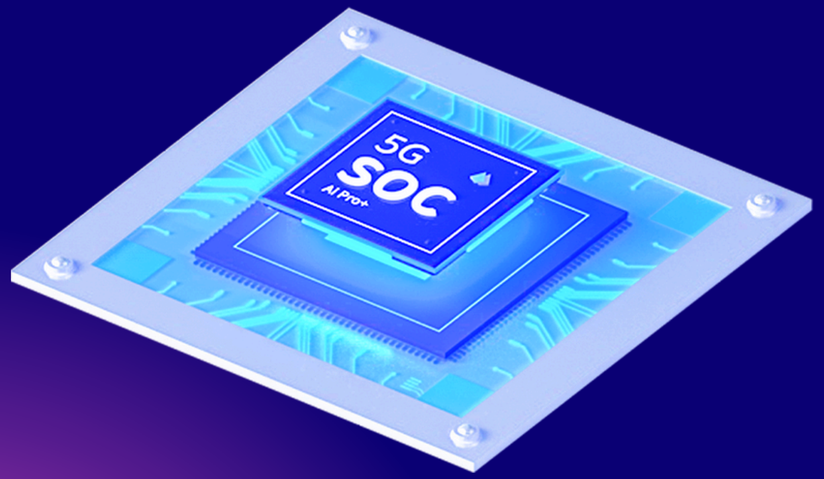
dimpybangoriya@gmail.com

ABOUT
THE PROJECT

This project demonstrates how to efficiently query data stored in Amazon S3 using Amazon Athena, with the help of AWS Glue Crawler for automatic schema discovery and cataloging. By leveraging AWS Glue, the process of structuring raw data into a queryable format becomes seamless, eliminating the need for manual schema definition. The project covers setting up an S3 bucket, configuring an AWS Glue Crawler, and executing SQL queries in Athena to analyze data directly from S3. This approach provides a cost-effective, serverless, and scalable solution for performing complex data analysis without the overhead of managing traditional databases.



INTRODUCTION



This case study showcases the process of querying data stored in Amazon S3 using Amazon Athena. We will utilize AWS Glue Crawler to automatically detect the data schema and make it accessible for querying in Athena. Furthermore, we will set up Athena to save query results in an S3 bucket.

PREREQUISITES

Before getting started, make sure you have the following prerequisites:

- An AWS account with the required permissions to create and manage S3 buckets, AWS Glue, and Amazon Athena.
- An S3 bucket containing the data files you intend to query.
- Access to the AWS CLI or AWS Management Console.

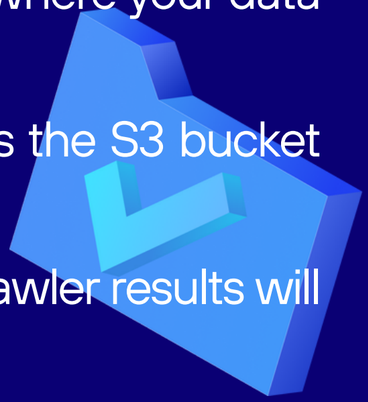
STEP-BY-STEP GUIDE

Step 1: Set up an S3 Bucket

1. Open the Amazon S3 console.
2. Create a new S3 bucket or use an existing one.
3. Upload your data files to S3 bucket.

Step 2: Create and Configure AWS Glue Crawler

1. Open the AWS Glue console.
2 Navigate to the Crawlers section and click Add Crawler.
3 Enter a name for your crawler and click Next.
4 Define the data store by selecting S3 and specifying the S3 bucket path where your data files are stored. Click Next.
5 Choose or create an IAM role with the necessary permissions to access the S3 bucket and AWS Glue.
6 Select or create a database in the AWS Glue Data Catalog where the crawler results will be stored.
7 Review the crawler configuration and click Finish.
8 Start the crawler to analyze the data and populate the schema in the Data Catalog.



Step 3: Configure Query Result Location in Athena

1. Open the Amazon Athena console.
2. Go to Settings and set the query result location to an S3 bucket where you want to store the results.
3. Save the settings.

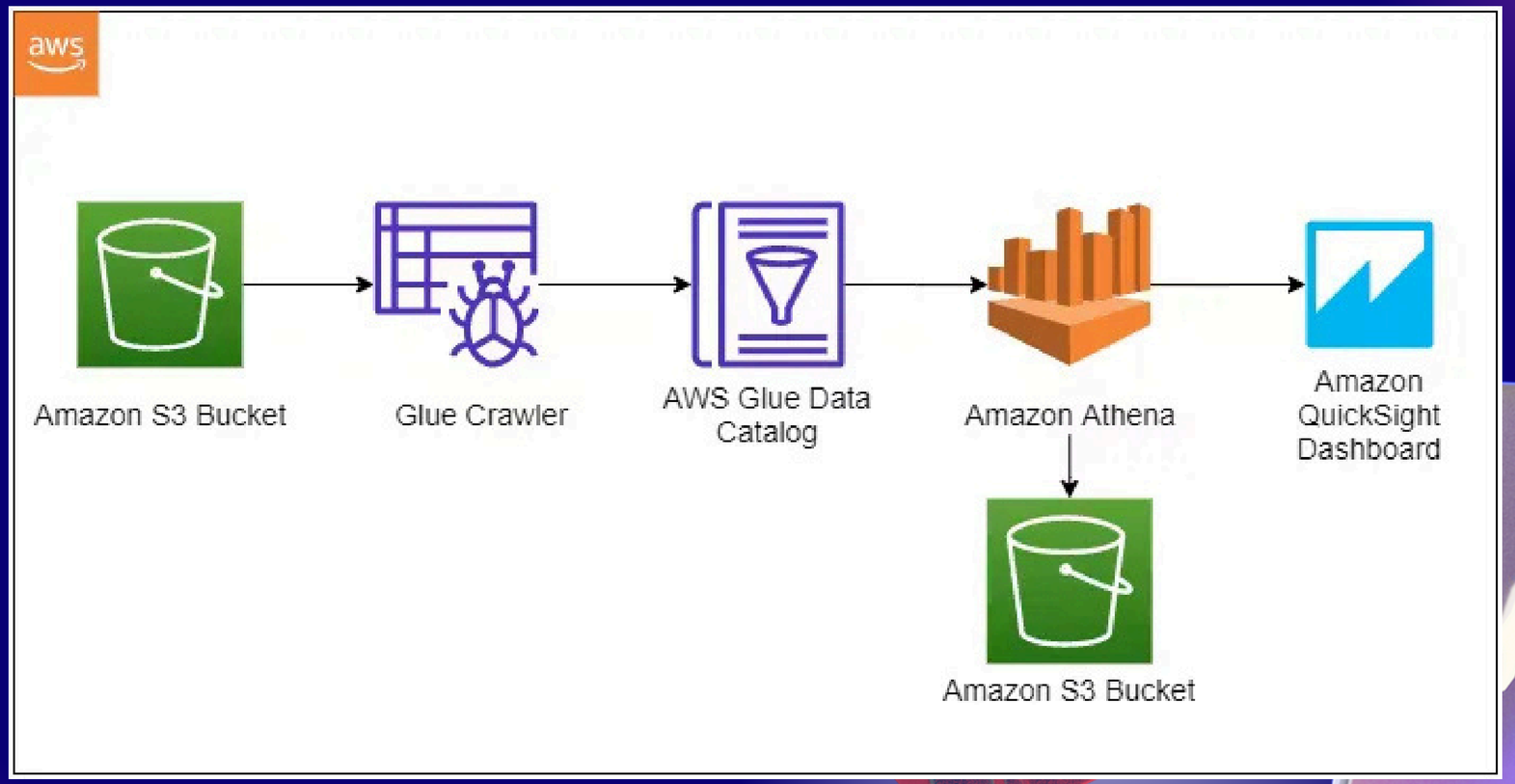
Step 4: Query Data Using Athena

1. In the Athena console, select the database created by the AWS Glue crawler.
2. Write an SQL query to retrieve data stored in S3. Example:

Query

```
SELECT * FROM your_table_name LIMIT 10;
```

3. Click Run Query to execute the SQL query.
4. View the results in the Athena console or check the stored query results in the specified S3 bucket.


aws
Amazon S3 Bucket
Glue Crawler
AWS Glue Data Catalog
Amazon Athena
Amazon QuickSight Dashboard
Amazon S3 Bucket

CONCLUSION

This project successfully demonstrated how to efficiently query data stored in Amazon S3 using Amazon Athena, with AWS Glue Crawler automating the schema discovery and cataloging process. By leveraging AWS Glue, we eliminated the need for manual schema definitions, making data readily accessible for analysis in Athena.

The implementation of this workflow enables serverless, cost-effective, and scalable querying, allowing users to extract valuable insights without managing a traditional database infrastructure. This approach is particularly beneficial for big data analytics, log analysis, and business intelligence use cases. Overall, the project highlights the efficiency of AWS services in simplifying data management and improving analytical capabilities.

THANK YOU

DIMPYBANGORIYA@GMAIL.COM